Reference wind farm selection for regional wind power prediction models

Nils Siebert George Kariniotakis
Center for Energy and Processes
Ecole des Mines de Paris, France
Phone: +33-4.93.95.75.01
nils.siebert@ensmp.fr, georges.kariniotakis@ensmp.fr

Abstract

Short-term wind power forecasting is recognized today as a major requirement for a secure and economic integration of wind generation in power systems. This paper deals with the case of regional forecasting of wind power with a large number of wind farms involved. Due to the large amount of potentially available information and also because part of the wind farms may not be "observable", forecasting systems use input from selected “reference” wind farms to predict the total wind power. The paper studies the influence of the reference farms on the prediction accuracy and proposes a methodology for their selection, based on advanced statistical analysis of the spatial-temporal characteristics of wind generation.
Keywords: regional forecasting, upscaling, reference farm selection, information, clustering

1 Introduction

Short-term wind power forecasting up to 48 or 72 hours ahead is recognized today by end users such as wind farm owners and power system operators as a major requirement for a secure and economic large-scale integration of wind generation. In the case of system operators forecasts of the expected power output of all turbines installed within a region are needed for various functions such as congestion management, reserves estimation, exchanges with neighboring systems etc. Regional wind power forecasting is thus defined as the prediction of the aggregated power output of wind farms spread over a geographical region.

Wind power prediction models may potentially use past measurements of power and meteorological variables as well as Numerical Weather Predictions (NWPs) of several variables (i.e. speed direction, temperature etc). In the frame of regional forecasting, and depending on the number of wind farms involved, this may result to a wealth of explanatory variables. Especially in the case of statistical forecasting approaches it becomes necessary to select among all available input variables those that are most relevant to the final objective. At a primary level the problem of variables selection can be simplified to a problem of wind farms selection.

In this paper a study is conducted to evaluate the impact of input selection on regional forecasting model performance, and several input selection methods that can help in model setup are examined.
The results of the proposed methodology are evaluated on a Danish case study of regional forecasting using a non-linear prediction model.

2 Regional wind power forecasting

The main characteristic of regional production is that the considered installed capacity is dispersed over a wide geographical area. This dispersion has one main advantage, which is the smoothing of the aggregated output of the wind farms. By summing the output of several wind farms, individual power variations are compensated and the resulting production presents much slower variations. This statistical smoothing effect leads to the main characteristic of regional forecasting which is that geographically dispersed production can be more accurately predicted than that of single wind farms [1].

At a regional or national scale, the ideal case would be to predict the output of each single wind farm in order to obtain the total generation in the next hours. However, wind power forecasting models require input data concerning the wind farms to compute the forecasts. This information can be of two natures, static and dynamic. Static information is the information necessary to describe a wind farm: the rated power, the power curve of the turbines, etc. This information is generally considered as not evolving over time. Dynamic information is time-series data that evolves over time such as production measurements and NWP for the wind farm.

In systems with an important number of wind farms, predicting the output of each single wind farm may be impossible due to a lack of static and\or dynamic data. For instance, on-line information is

usually not available for all wind farms. This is because Supervisory Control and Data Acquisition (SCADA) systems may not be installed in all farms (grid codes in some countries impose such infrastructure above a certain level of installed farm capacity).

The absence of SCADA systems means that the actual total wind production is seldom known accurately. However, even if the actual power each wind park injects into the grid at a given time can be known with some delay, the actual available capacity at that time in the region is hard to know. This can be particularly hampering for statistical model training since the measured regional power cannot be corrected for unavailable capacity, which leads to noise in the training data. However, it can be assumed that statistical smoothing also acts upon the available capacity, which therefore does not vary significantly over time.

Because of the usual lack of static and especially dynamic information, the mainstream approach to regional prediction is to use upscaling methods. The idea behind these methods is to use the available static and dynamic information to extrapolate the best possible forecast of the aggregated power output of all the farms in one area. In this paper, the wind farms for which dynamic data is available will be referred to as reference wind farms.

3 State of the art

Short-term wind power forecasting has been an active research field and numerous approaches have been proposed [2, 3]. In most cases the proposed methods have been designed for single wind farm forecasting. The models developed so far can be classified as using either a physical or a statistical approach. In some models, a combination of both is used, as indeed both approaches can be needed for successful forecasts.

The physical models use NWPs and physical considerations to reach the best possible estimate of the local wind speed before using a power curve to convert the wind forecast into the power forecast.

Statistical models try to find the relationships between a wealth of explanatory variables including NWPs, and online measured power data. Often, black-box models like advanced Recursive Least Squares or Artificial Neural Networks are used. Some statistical approaches actually employ gray-box models, where some knowledge of the wind power properties is used to tune the models to the specific domain.

Up to now several upscaling approaches have been developed mainly for application in countries with large-scale wind integration (i.e. Denmark, Germany). The algorithm proposed in [4, 5] uses a physical model that takes into account the site description (i.e. hub height, terrain description) to provide forecasts for reference farms, which are then used by an upscaling algorithm to provide the regional power. Reference farms are chosen according to their physical characteristics as being representative of the farms in the region.

Another regional prediction model is RegioPred [6]. This model is based on the single wind farm prediction model LocalPred, which uses high-resolution Computational Fluid Dynamics (CFD) models and statistical power curve modeling to compute the power forecasts. The reference wind farms used in this model are selected using cluster analysis.

In [7], an upscaling algorithm based on two artificial neural networks is used to provide regional forecasts. The first neural network is used to compute the power output of a certain number of reference wind farms. These forecasts are then used as input to a second network that computes the forecasts of the total regional production.

The approach described in [8] uses conditional parametric models to compute the regional forecasts. In this model a two-branch approach is used. The first branch uses models to produce single wind farm forecasts for the reference wind farms. These forecasts are then used to compute forecasts of the power generated in subregions. Finally, these forecasts are added to provide a forecast for the total regional production. A second branch of the model uses the data from reference farms to directly produce sub-regional forecasts, which are then aggregated to produce a second regional forecast. The regional forecasts computed by each branch are then combined through weighted averaging to obtain the final total regional forecast.

In [9] Fuzzy-Neural Networks (F-NN) are used to compute regional wind power. Several model configurations are tested. These range from upscaling using data from a single reference farm to a cluster forecasting approach, where predictions are made for sub-regions, and the forecasts for each sub-region are then aggregated to provide the regional forecast. The reference wind farms were selected based on a correlation analysis of their power production with the regional power production.

As mentioned above, the main problem in regional forecasting is that dynamic information for all wind farms is not usually available. However, the number of wind farms for which such data is available can still be quite important. Furthermore, explanatory input may increase depending on the available NWPs (wind speed and direction for several altitudes, temperature, pressure, humidity, etc). For a case where many reference wind farms are available, up to a hundred or more variables may be available to the model.

Since upscaling models are statistical, they are subject to certain constraints, the first being the number of explanatory variables taken into account. As the number of variables increases, the number of

parameters in the model increases. This is against the principle of "parsimony" in prediction models while it can also lead to higher uncertainty on the estimates of these parameters [10]. This later depends on the number of data available for the estimation of the models parameters. Over-dimensioned models may lead to overfitting of the models parameters and to low generalization (out-of-sample) performance.

Another aspect to take into account is the redundancy of the available variables, especially those provided by the NWP models. For example, over a region, the wind speed forecasts provided for two neighboring wind farms will usually present high correlations. Using these two forecasts will not necessarily provide the upscaling model with more information than that obtained by using only one of the forecasts. Although redundancy in the variables may allow some noise reduction [11], using too many redundant variables may lead to a reduction of model's performance [12].

Given the risks of overfitting and of redundancy in the available explanatory variables, choosing which variables to use is a crucial aspect of the upscaling modeling problem.

For these reasons investigating the impact of reference wind farm selection on the accuracy of regional and upscaling forecasting models is necessary. It is of interest to analyze what the characteristics of the reference wind farms that lead to the best forecast accuracy are. A corollary to this question is then how can the best reference wind farm combination be determined, which relates to the problem of input selection for a model when a multitude of explanatory variables are available. In the following sections, a methodology is proposed to answer these questions.

4 Reference wind farm selection study

4.1 Proposed method

To investigate the impact of reference wind farm selection on the performance of upscaling models, the test case of the Jutland-Funen area in Denmark (see section 4.2.1) was considered. Data from 23 wind farms (~10% of total capacity) as well as time-series of the total power are available. Potentially the number of reference farms can be the total of 23 (denoted as n). At a first stage, all alternative upscaling schemes resulting from the 2^n-1 possible combinations of reference wind farms were evaluated. The assessment of all these schemes required a computationally efficient upscaling model; it would have been unfeasible to use CPU-expensive advanced models. For this, the Regressive Power Curve (RPC) model, presented below, was developed. This model is computationally efficient and its performance was found to be similar to that of more advanced models. The hypothesis made here is that the performance of a reference wind farm combination is more influenced by the information carried in the input data rather than the type of prediction model. Thus, some insight on the influence of reference farms can be gained which may allow the development of a rather generic reference wind farm selection procedure.

4.1.1 The Regressive Power Curve (RPC) model

The RPC model is based on a transformation of the NWP wind speed forecast values into power using the characteristic curve of the wind farm. The simplest way of doing this is to use the wind turbines manufacturer's power curve and the wind speed forecast extrapolated to the hub height using the logarithmic profile of the wind. Although simple, this approach may provide reasonably good forecasts. However, the sum-up of the wind turbines curves is not necessarily a good model for the wind farm. On the other hand, in a physical approach where the NWPs are provided for a grid of points around the wind farm, a downscaling procedure should be applied to have reliable wind speed forecasts at the level of the wind farm. This would require a significant amount of static information on the wind farm (i.e. roughness of the terrain), which might no be readily available.

In order to obtain a model that can reduce the above-mentioned approximations and whose only information requirements are the time-series under consideration, a statistical approach was preferred. The aim here is to model the relationship between the wind speed forecasts and the power output of the wind farm without any other considerations. This approach is often referred as "power curve modeling". To model the power curve *f(ws)*, a piecewise least squares linear fitting of the wind-speed to power relation is proposed. To account for the horizon for which the prediction is being computed a separate power curve is defined for each forecast horizon. In this way, variations in the NWP performance for different horizons can be captured and corrected to some extent. The obtained power curve can be then written as a function of the wind speed and horizon as:

$$f_k\left(ws\left(t+k|t\right)\right)=\begin{cases}\alpha_{1,k}ws\left(t+k|t\right)+\beta_{1,k} & \text{if } \gamma_{0,k}\le ws\left(t+k|t\right)<\gamma_{1,k}\\ \vdots & \\ \alpha_{i,k}ws\left(t+k|t\right)+\beta_{i,k} & \text{if } \gamma_{i-1,k}\le ws\left(t+k|t\right)<\gamma_{i,k}\\ \vdots & \\ \alpha_{m,k}ws\left(t+k|t\right)+\beta_{m,k} & \text{if } \gamma_{m-1,k}\le ws\left(t+k|t\right)\le\gamma_{m,k}\end{cases} \quad (1)$$

where:

- $f_k\left(ws\left(t+k|t\right)\right)$ is the power associated to the wind speed forecast $ws\left(t+k|t\right)$,
- $\alpha_{i,k}$ and $\beta_{i,k}$ are the parameters of the local linear approximation of the power curve corresponding to horizon k.

- $[\gamma_{i-1,k},\gamma_{i,k}[$ the m bins over which the least squares fitting is performed.

The above model is the basis for the RPC model proposed. From (1) it is clear that the size of the bins used to compute the local least squares approximations will have an impact on the performance of the model. Very small bins will capture the noise in the data and very large bins will over-smooth the power curve.

It is known that if online power measures are available through a SCADA they can contribute to a better performance of a prediction model in the short-term (i.e. 0-6 hours). To account for this, a multiple regression between the power curve model output and the last available power measure is applied. In this way, the output for each forecast horizon k can be written as:

$$F_k(ws(t+k|t),p(t)) = a_k f_k(ws(t+k|t)) + b_k p(t) + c_k \quad (2)$$

where:

- $F_k(.)$ is the RPC model overall function,
- $p(t)$ is the power measure at time t,
- f_k is the power curve computed for horizon k,
- a_k, b_k, and c_k are the multiple regression coefficients.

To determine the best bin sizes bs_k^* in (1) and the values of the $\alpha_{i,k}$, $\beta_{i,k}$, a_k, b_k and c_k parameters, an iterative procedure is used (Algorithm 1). To apply this procedure the dataset is divided into three subsets: a learning set, a validation set and a testing set. The steps described in Algorithm 1 constitute the learning phase for this model.

For bin size $bs = size_1$ to $size_l$ do
  Compute the $\alpha_{i,k}$, $\beta_{i,k}$, a_k, b_k and c_k parameters on the learning set.
  Using these parameters, compute the sum of square errors for each horizon on the validation set.
  Store the performance of each bin size for each horizon.
End for
For each horizon select the bin size which led to the smallest errors.
For each horizon
  Compute the $\alpha_{i,k}$, $\beta_{i,k}$, a_k, b_k and c_k parameters on the learning and validation sets using the best bin size.
End for

Algorithm 1: RPC model's learning algorithm.

Once the model parameter values have been estimated, the model performance can be evaluated on the testing dataset.

The RPC model can be used for both single wind farm forecasting and upscaling purposes. For upscaling, the following approach is followed: for a given reference wind farm combination, the data from each wind farm is used to compute a forecast of that wind farm's production. The individual wind farm forecasts are then linearly combined to provide the final regional forecast.

4.1.2 Evaluation of the RPC model

To validate this model as a good candidate for evaluation purposes, single wind farm forecasting was first examined. The benchmarking was done on the datasets used in [13]. More specifically, data were used from the Tunø Knob, Golagh, and Alaiz wind farms in Denmark, Ireland and Spain respectively.

In the following figures the normalized mean absolute error (NMAE) for each horizon of the F-NN [14] and the RPC models are compared.

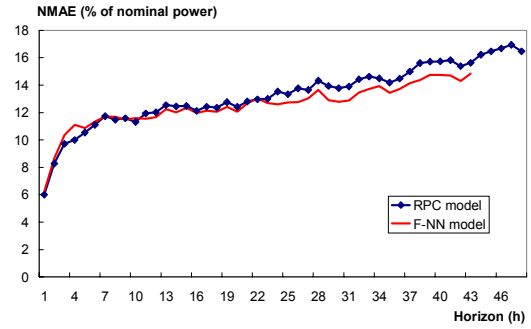

Figure 1: NMAE as a function of the forecast horizon of the F-NN and RPC models for the Tunø Knob wind farm.

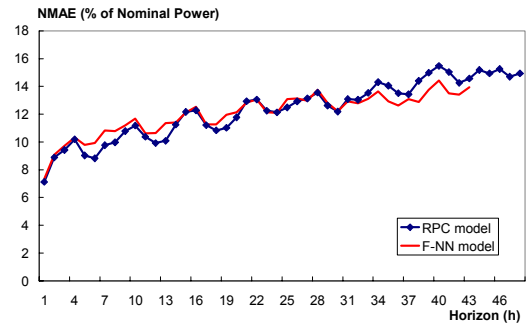

Figure 2: NMAE as a function of the forecast horizon of the F-NN and RPC models for the Golagh wind farm.

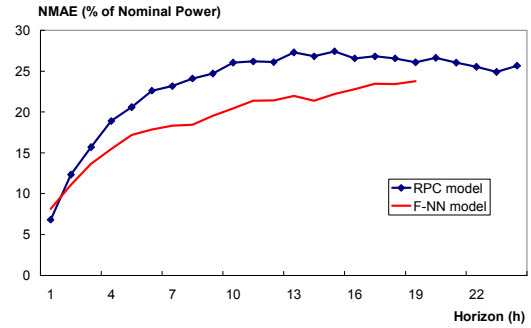

Figure 3: NMAE as a function of the forecast horizon of the F-NN and RPC models for the Alaiz wind farm.

These results show that the RPC model performs quite well when compared to a more advanced one. For farms located in offshore or complex terrain like Tunø Knob and Golagh, the performance of the two models is very close. This can be explained by the fact that for simpler terrain types the NWP models provide better estimates of the wind, and the relations between input and output variables are less complex.

For wind farms in very complex terrain like Alaiz, the F-NN model outperforms the simpler RPC model. The complexity of the terrain and the ensuing local effects are not well captured by the NWP models. The F-NN is better suited to these complex cases because it can more precisely model the intricate relations between input and output variables.

The performance of the RPC model was also evaluated on the Danish dataset. For this evaluation, the RPC model was used as a direct upscaling model using the averaged NWP data and the aggregated power data of the 23 wind farms as online measurement data. This is the same configuration used for the F-NN model on the Danish case. The comparison of both models is shown in Figure 4.

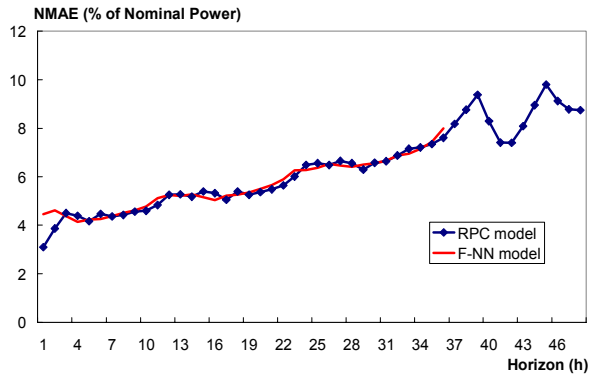

Figure 4: NMAE over all forecast horizons of the F-NN and RPC direct upscaling configuration using averaged NWP data for the Jutland-Funen area.

For this case the performance of the RPC model compares very favorably with the performance of the F-NN model. The difference in performance for the first two hours is due to the fact that the F-NN architecture used for this comparison does not take into account production measurements.

This benchmarking is satisfactory in that the performance of the RPC model closely follows the performance of a proven advanced model. Furthermore, the CPU time needed to run this model for a case study is in the order of a fraction of a minute. These two properties make this model a good candidate for reference wind farm selection methodology. Also, in the frame of regional forecasting, where tens of wind farms are available as potential reference farms, this model can serve as an exploratory tool to evaluate the predictability of each wind farm.

4.2 Results

4.2.1 Case study

The dataset consists of measured power output for 23 wind farms as well as of the total power output of all wind farms located in the Jutland-Funen area (more than 2 GW installed capacity). The data cover a period spanning from the 1st of January 2003 to the 31st of July 2004.

NWP data from Hirlam is available for all 23 wind farms. The NWP runs are provided twice a day: at 12h00 and 00h00. Forecasts cover the NWP run time and the following 48 hours, with an hourly resolution.

In the present study the online power measurements were not considered in order to focus exclusively on the impact of NWP data selection, therefore only 10-meter wind speed predictions were considered as input to the RPC model.

4.2.2 Results

In this section the results of the wind farm combination study are presented. Given the number of tested combinations, only the results of the best combinations for each cardinality are presented.

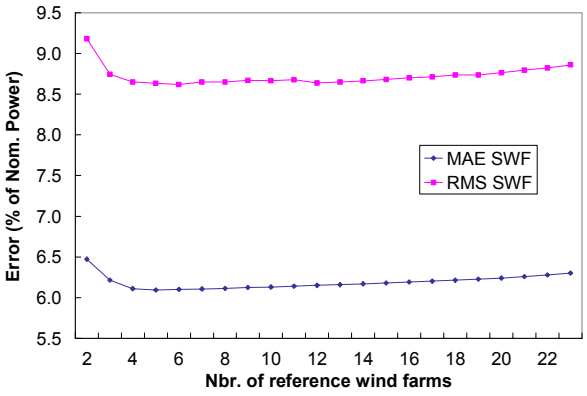

Figure 5: Optimal performance in terms of NMAE and NRMSE for various reference wind farm combinations.

As can be expected, increasing the number of reference wind farms initially increases the performance of the upscaling model. However, there appears to be a point after which adding reference wind farms actually reduces the forecasting model's performance. This can be explained by the fact that increasing the number of reference farms increases the amount of information available to the model, but also the amount of noise the model has to filter. Since the explanatory variables provided by the reference wind farms are correlated, the amount of additional information a wind farm provides decreases with the number of selected wind farms. For this reason, after a certain number of wind farms are selected, the additional information provided by an extra farm is outweighed by the additional noise it adds to the model's input.

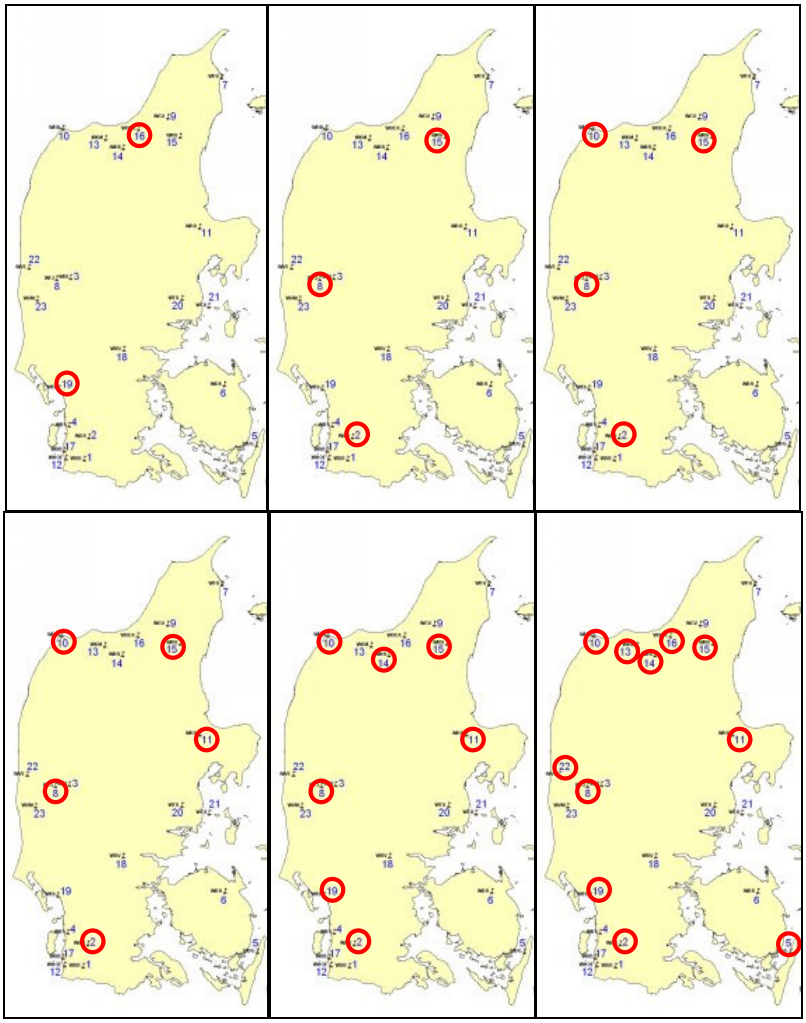

Figure 6: Geographic position of optimally chosen reference wind farms for combination of cardinality: 2, 3 ,4 ,5 ,7 and 11.

To better understand the parameters that characterize the farms belonging to the best combinations, their size, the correlation of their production to the regional production, and the correlation of their NWP forecasts with that of other farms were examined. No clear relation between these parameters and combination performance could be established. However, by examining the spatial distribution of the farms some insight on the characteristics of a good combination can be gained.

As can be seen in Figure 6, the best combinations appear to be those that offer the best coverage of the region in terms of meteorological data. In this sense, the best two-farm combination (WF16 and WF19) provides the best possible information on the meteorological condition reigning over the area. When three-farm combinations are considered, the best solution is given by combination 2-8-15, which offers more detailed coverage. The evolution of selected wind farms as their number increases clearly shows a progressive coverage of the considered region.

5 Reference Wind Farm Selection Methodology

The problem of input variable selection for statistical regression models has been an area of intense research for many years. Many methods have been proposed to solve the optimization problem that can be defined as: given a regression model and a set of explanatory variables, find the subset of explanatory variables that leads to the least forecast error.

5.1 Investigation of alternative approaches

5.1.1 Clustering approach

Given the apparent influence of the geographical distribution of the best wind farm combinations, one approach that was tested in the frame of this work was to use the k-means clustering algorithm [15]. With this we classify the wind farms into *k* clusters. Within each

cluster, the farm closest to the cluster centroid is chosen as reference farm. The clustering algorithm is run n times (where n is the total number of available reference farms) to determine n reference farm combinations of cardinalities *1* to n. The forecasting model is then run for every selected wind farm combination to determine which combination is best.

With this approach, an important question is the metric chosen to compute the distance between wind farms. Two approaches were considered: the Euclidean distance using the wind farms' UTM coordinates and the correlation between the NWP wind speeds given for each farm.

5.1.2 Information theoretic approach

The second approach is based on the principle of entropy. Shannon's entropy of a discreet random variable X is defined by:

$$H(X)=-\sum_{i=1}^{m} p_X(a_i)\ln(p_X(a_i)), \quad (2)$$

where $p_X(a_i)=P(X=a_i)$ is the probability mass function of variable X, and $\{a_1, \ldots, a_m\}$ the possible outcomes of the variable. For continuous variables, the sum becomes an integral and the probability mass functions are replaced by probability density functions.

The conditional entropy H of two variables X and Y is defined by:

$$H(X\,|\,Y)=H(X,Y)-H(Y). \quad (3)$$

The mutual information between two discreet variables is defined as:

$$I(X;Y)=I(Y;X)=H(X)-H(X\,|\,Y), \quad (4)$$

which is equal to:

$$I(X;Y)=\sum_{i=1}^{m_X}\sum_{j=1}^{m_Y} p_{X,Y}(a_i,b_j)\ln\frac{p_{X,Y}(a_i,b_j)}{p_X(a_i)p_Y(b_j)}. \quad (5)$$

In case of continuous variables the double sum becomes a double integral.

Entropy can be seen as the uncertainty concerning the outcome of a random variable while the mutual information can be seen as the reduction of uncertainty concerning the outcome due to the knowledge of the outcome of another variable. Therefore, a model is expected to produce better forecasts if it uses as input the explanatory variables that have the highest mutual information with the dependant variable.

The explanatory variable can be a multivariate random variable whose components are the explanatory variables. In this case, the problem of input selection for a model can be described as finding a subset S of k variables from a set F of n available explanatory variables which maximizes *I(X;Y)*, where X is the dependant variable and Y the multivariate random variable composed by the explanatory variables in S.

Solving such a problem is not straightforward. To compute *I(X;Y)*, where Y is a continuous multivariate variable requires the estimation of multivariate distributions. In order to obtain an accurate estimate of multivariate distributions for dimensions higher than 4 or 5, datasets with more than one hundred thousand samples are necessary [16]. Clearly, such large datasets are not readily available in the field of wind power forecasting. A further problem is that this approach would require computing the mutual information for all C_n^k explanatory variable combinations.

To overcome these obstacles, Battiti [17] proposed the MIFS (mutual information based feature selection) greedy algorithm. This algorithm is based on the evaluation of the mutual information between the dependent variable and each explanatory variable, and on the evaluation of the mutual information between single explanatory variables. The algorithm is summarized as:

Define set F of all explanatory variables.
Define set S of all selected variables.
Find variable $Y\in F$ which maximizes *I(X;Y)*, where X is the dependent variable
Remove Y from F and add it to S.
Repeat until $|S|=k$ the desired number of selected variables.
  Find $Y\in F$ which maximizes $I(X;Y)-\beta\sum_{s\in S} I(Y,s)$
  Remove Y from F and add it to S
Output the set S which contains the selected variables.

Algorithm 2: MIFS algorithm.

We propose to use this algorithm to determine the farm combinations for all cardinalities. After the combinations are selected, the forecasts are computed with the upscaling model for each combination in order to find the best one.

5.1.3 RPC-based wrapper wind farm selection method

When considering reference wind farm combinations the total number of possible combinations is equal to 2^n-1, where n are the available reference wind farms; for the Danish case this translates to more than 8 million combinations. Evaluating all possible combinations would therefore not be reasonably feasible for cases with many potential reference farms.

To overcome this difficulty, the RPC model can be used in a greedy forward selection algorithm to determine the best wind farm combination to use. This approach is described below:

Evaluate all 2-farm combinations
For i=3 to n do
  Determine SC_{i-1} the m_{wf} best combinations of cardinality i-1.
  Evaluate the farm combinations derived from combinations in SC_{i-1}.
End for
Select the best combination from those computed.

Algorithm 3: RPC based selection algorithm.

where SC_i the subset of m_{wf} best combinations of cardinality i.

From one step to the next in this algorithm, the best combinations of step i-1 are used as the basis for those that will be evaluated in step i. For example if there are 5 reference wind farms and the two best 2-farm combinations are: 1-2, 1-3, the 3-farm combinations that will be computed are:

Best 2-farm combinations	1-2	1-3
Examined 3-farm combinations	1-2-3	1-3-4
	1-2-4	1-3-5
	1-2-5	

Table 1: Example of combination derivation from an SC_2 subset.

Through the appropriate setting of m_{wf}, this wrapper method permits to reduce the number of combinations evaluated while retaining a wide enough search scope. The choice of the m_{wf} parameter depends on the transition stability of the optimal combinations from one cardinality to the next. By transition stability we mean that from one cardinality to the next the optimal combinations share an important number of reference farms. For example, in a ten-reference wind farm case, if the optimal four-farm combination is 1-2-3-4 and the optimal five-farm combination is 1-2-3-4-8, the transition can be said to be stable since the five-farm combination is derived from the four-farm combination by the adjunction of farm 8. However if the optimal five-farm combination is 5-6-7-8-9, the transition is unstable since the optimal five-farm combination cannot be derived from the optimal four-farm combination. If transitions are very stable, i.e. the best combinations from one cardinality are derived from the best combinations of inferior cardinality, the m_{wf} parameter can be set to a low value. However, if transitions are unstable, setting the m_{wf} parameter to higher value will be necessary in order to reduce the risk of missing the optimal combination.

5.2 Results

5.2.1 Clustering approach

The tables below provide the combinations found with the k-means clustering approach.

Computed combination	Mean NMAE (% of Nom. Power) over all horizons for computed combination	Mean NMAE (% of Nom. Power) over all horizons for best combination of same cardinality
16-18	6.59	6.45
6-16-19	6.47	6.18
4-6-8-16	6.25	6.14
4-6-8-13-15	6.24	6.13
4-5-8-9-13-20	6.28	6.14

Table 2: Combination results from the clustering approach with the eucliddian metric and the farms' UTM coordinates.

Computed combination	Mean NMAE (% of Nom. Power) over all horizons for computed combination	Mean NMAE (% of Nom. Power) over all horizons for best combination of same cardinality
16-21	7.98	6.45
2-16-21	6.59	6.18
2-6-16-22	6.34	6.14
2-5-6-8-16	6.22	6.13
2-5-6-8-11-15	6.23	6.14

Table 3: Combination results from the clustering approach with the correlation metric and the predicted wind speed for the farms.

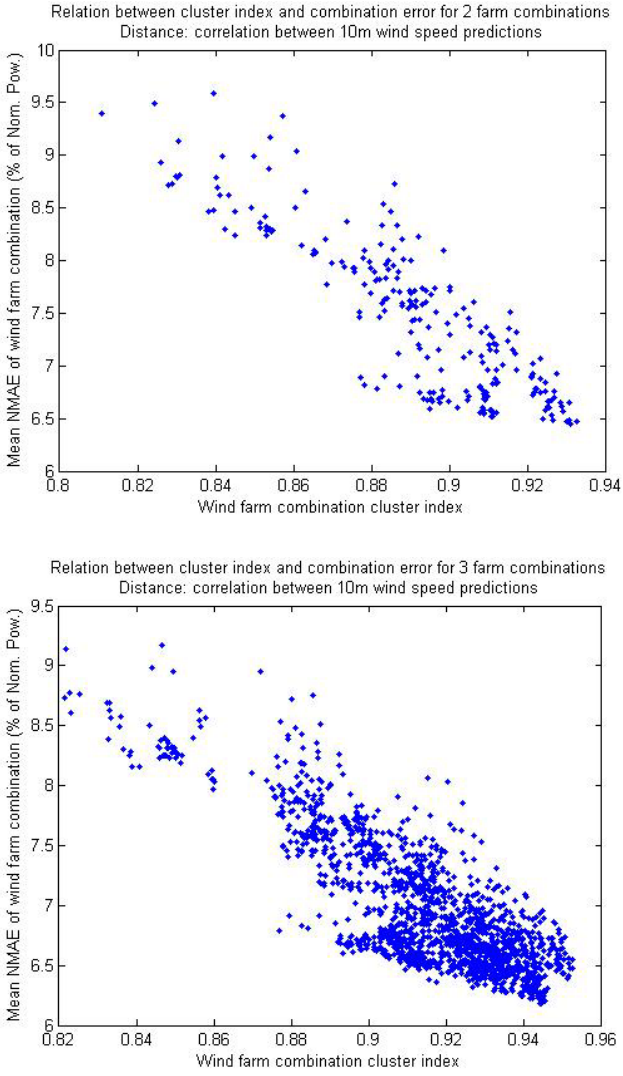

Figure 7: Relation between wind farm combination cluster index and mean NMAE obtained for the given wind farm combinations.

Given the relatively poor correspondence between the best combinations and the combinations found using the clustering approaches, we chose to examine the behavior of the clusters more closely. To do so, the farms were clustered using a defined combination of wind farms as cluster centroids. Once all farms have

been assigned to the clusters based on their distance to that cluster's centroid (one of the reference farms of the combination under scrutiny), a cluster index is computed for that combination by summing the average within cluster distance to the centroid. This allows attributing a score to every wind farm combination based on the goodness of the clustering achieved by using the farms of that combination as centroids.

Figure 7 depicts the results of this procedure where the distance is the correlation of wind speed forecasts between wind farms. The relation between the wind farm combination cluster index and the mean NMAE over all horizons for each wind farm combination obtained using the RPC model with NWP data as the only input. The results show a clear relation between the cluster index values and the performance of the wind farm combinations. This can be a useful indicator of which wind farm combinations are best. However, from these figures it appears that there are two behaviors concerning the wind farm combinations. The combinations are grouped according to two linear relations between the cluster index value and the mean NMAE. This behavior can be seen for the 2-farm combination and even more clearly for the 3-farm combination case.

The explanation for this behavior is not clear; however it could possibly be linked to the installed capacity represented by the reference farms. The RPC model might be able to capture such relations from the data whereas an approach based on the sole study of the relation between wind farm variables might not.

5.2.2 Information theoretic approach

Computed combination	Mean NMAE (% of Nom. Power) over all horizons for computed combination	Mean NMAE (% of Nom. Power) over all horizons for best combination of same cardinality
2-3	6.67	6.45
2-3-9	6.27	6.18
2-3-9-21-	6.3	6.14
2-3-9-17-21	6.38	6.13
2-3-9-10-17-21	6.27	6.14

Table 4: Combination results found using the MIFS algorithm.

By using Battiti's algorithm relatively good combinations can be found. However, the fact that the algorithm is greedy, i.e. it keeps the best solution at each step, it cannot necessarily capture the "instability" of optimal wind farm combinations that exists for low combination cardinalities. From the results presented earlier, it is clear that there is an important difference between the optimal 2-farm combination and the optimal 3-farm combination. However, given that the algorithm only extends the 2-farm combination to a 3-farm combination, it necessarily fails to capture this important modification that conditions the evolution of the optimal farm combinations for higher cardinalities.

5.2.3 RPC based wrapper wind farm selection method

It can be noticed that these architectures are relatively "stable" in the sense that a given architecture is similar to the one of inferior cardinality, especially for the higher cardinalities.

As seen in section 5.1.3, the value given to the m_{wf} parameter is highly dependent on the stability of transitions from one combination cardinality to the next. In a first stage all combinations were computed, the wrapper method was then run using different values for m_{wf}. With values of m_{wf}=20 the best combinations for each step were found and computation time remained reasonable. However, with this method the number of computed combinations remains important. For this case study, several thousand combinations were computed, which nonetheless compares favorably with the more than 8 million possible combinations.

6 Conclusions

In this paper we have shown that input variable selection can have an important impact on the accuracy of upscaling models. The redundancy of NWP variables and the inherent limit of statistical models with respect to the number of input variables should be considered when setting up an upscaling module.

Several input selection methods were tested. None of the filter methods gave optimal results but they do allow choosing combinations that lead to better performance than that which can be expected from a randomly chosen combination. In cases where model set up time is short these approaches can prove useful.

In cases where CPU time is not an issue for model set-up, using a simplified model in a greedy wrapper approach as proposed in this paper, can permit finding an optimal solution without explicitly evaluating every combination.

7 Acknowledgements

This work was performed in the frame of the ANEMOS Project (ENK5-CT-2002-00665) funded in part by the European Commission as well as in the frame of the ADEME Grant TEZ04-32 and Convention 0205076. Acknowledgments are due to ELSAM, ESB National Grid, and ACCIONA (ex. EHN) for providing the data for this work.

8 References

[1] Focken U. et al, “Short-term prediction of the aggregated power output of wind farms-A statistical analysis of the reduction of the prediction error by spatial smoothing effects”, *J. Wind Eng. Ind. Aerodyn.*, Vol. 90 (3), 2002, pp. 231-246.

[2] Giebel, G., Kariniotakis, G., Brownsword, R., "The State-of-the-Art in Short-Term Prediction of Wind Power - From a Danish Perspective", in *Proc. of the 4th International Workshop on Large-Scale Integration of Wind Power and Transmission Networks for Offshore Wind farms*, Billund, Denmark, October 20-21, 2003.

[3] Kariniotakis, G., Pinson, P., Siebert, N., Giebel, G., Bartelmie, R., “The State of the Art in Short-term Prediction of Wind Power - From an Offshore Perspective”, *Symposium ADEME, IFREMER*, "Renewable energies at sea", Brest, France, October 20-21, 2004.

[4] Focken U. et al, “Previento - A wind power prediction system with an innovative upscaling algorithm”, in *Proc. of the 2001 European Wind Energy Association Conference*, EWEC’01, Copenhagen, Denmark, pp. 826-829, 2-6 July 2001.

[5] Focken U., Lange M., Heinemann D., “Previento - Regional wind power prediction with risk control”, in *Proc. of the 2002 Global Windpower Conference*, Paris, France, 2-5 April 2002.

[6] Marti, I., *et al*, "LocalPred and RegioPred. Advanced tools for wind energy prediction in complex terrain", in *Proc. of the European Wind Energy Conference EWEC 2003*, Madrid, Spain, 16-19June, 2003.

[7] Ernst B., Rohrig K., Schorn P., Regber H., “Managing 3000 MW power in a transmission system operation centre”, in *Proc. of the 2001 European Wind Energy Association Conference*, EWEC’01, Copenhagen, Denmark, pp.890-893, 2-6 July 2001.

[8] Nielsen T. S., Madsen H., Nielsen H. Aa., Landberg L, Giebel G., “Prediction of regional wind power”, in *Proc. of the 2002 Global Windpower Conference*, Paris, France, 2-5 April 2002.

[9] Pinson P., Siebert N., Kariniotakis G., “Forecasting of Regional Wind Generation by a Dynamic Fuzzy-Neural Networks Based Upscaling Approach”, in *Proc. of the 2003 European Wind Energy Association Conference EWEC’03*, Madrid, Spain, 16-19 June 2003.

[10] Chatfield C., *Time-series Forecasting*, Chapman & Hall/CRC, London, 2000.

[11] Guyon I., Elisseeff A., “An Introduction to Variable and Feature Selection”, *Journal of Machine Learning Research*, Vol 3, pp. 1157-1182, 2003.

[12] Kohavi R., John G. H., “Wrappers for feature subset selection”, *Artificial Inteligence*, Vol. 97 (1-2), pp. 273-324, 1997.

[13] Kariniotakis G,. Marti I., *et al.*, “What Performance Can Be Expected by Short-term Wind Power Prediction Models Depending on Site Characteristics?”, in *Proceedings of the European Wind Energy Conference EWEC 2004*, London, UK, 22-25 November, 2004.

[14] Kariniotakis G., Mayer D., “An advanced on-line wind resource prediction system for the optimal management of wind parks”, in *Proceedings of the 2002 MedPower Conference*, Athens, Greece, 4-6 Nov. 2002.

[15] Marques de Sá J. P., *Pattern Recognition: Concepts, Methods ans Applications*, Springer Verlag, Berlin, 2001.

[16] Scott D. W., *Multivariate Density Estimation: Theory, Practice, and Visualization,* John Wiley & Sons, Inc., 1992.

[17] Battiti R., “Using Mutual Information for Selecting Features in Supervised Neural Net Learning”, *IEEE Trans. on Neural Networks*, Vol. 5, No. 3, July, 1994.